AUTOMATIC SPEECH SEGMENTATION AND VERIFICATION METHOD AND SYSTEM

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to the technical field of speech synthesis and, more particularly, to an automatic speech segmentation and verification method and system.

2. Description of Related Art

Currently, in the technical field of speech synthesis, the concatenative synthesis based on a large speech corpus has become a popular approach to speech synthesis system because of the high acoustical quality and natural prosody that is provided. The key issues of speech synthesis systems include the number of synthesis units, the quality of the recorded material, the decision and selection of synthesis unit types, and the generation of natural rhyme and conjunction of synthesis units. Due to improvements in computer processing abilities, and the ubiquity of high capacity hard discs, the prior art speech synthesis system stores thousands of synthesis units to find proper synthesis units.

In the prior art speech synthesis method that uses large speech corpus, a main source of the synthesis units is a predetermined recording script, which is recorded by professional technicians with professional recording equipment. The computer system then automatically segments a recorded file according to phonetic information in the recording script to abstract speech units for the speech synthesis system.

However, sometimes the prior art segmentation position is incorrect, and a huge recording script requires performance and recording times, since even a professional technician can make pronunciation errors, insertion errors, and deletion errors, or co-articulation due to excessively quick speech. Since the segmentation positional accuracy and synthesis units' quality and correction directly affects the output speech synthesis quality, it is very important to improve the confidence measure to sieve bad recording material and redo it.

In addition to checking for segmentation errors, human effort is also required to check the consistency between the recorded speech and the phonetic transcription of the text script that was supposed to be read during recording. However, manual detection is laborious and is easily affected by personal reasons, which can cause totally different opinions about the same recording material instead of providing a consistent objective standard.

There are a lot of prior arts in this speech verification technique field, such as U.S. patent No. 6292778, which uses a speech recognizer and a task-independent utterance verifier to improve the word/phrase/sentence verification ratio. In this technique, the utterance verifier employs subword and anti-subword models to produce the first and second likelihoods for each recognized subword in the input speech. The utterance verifier determines a subword verification score as the log ratio of the first and second likelihoods, and the utterance verifier combines the subword verification scores to produce a word/phrase/sentence verification score, and compares that score to a predetermined threshold. U.S. patent No. 6125345 uses a speech

recognizer to generate one or more confidence measures; the recognized speech output by the speech recognizer is input to the recognition verifier, which outputs one or more confidence measures. The confidence measures output by the speech recognizer and the recognition verifier are normalized and then input to an integrator. The integrator uses a multi-layer perceptron (MLP) to integrate the various confidence measures from the speech recognizer and the recognition verifier, and then determines whether the recognized utterance hypothesis generated by the speech recognizer should be accepted or rejected. U.S. patent No. 5675706 uses subword level verification and string level verification to determine whether an unknown input speech does indeed contain the recognized keyword, or consists of speech or other sounds that do not contain any of the predetermined recognizer keywords. The subword level verification stage verifies each subword segment in the input speech as determined by a Hidden Markov Model recognizer to determine if that segment consists of the sound corresponding to the subword that the HMM recognizer assigned to that segment. The string level verification stage combines the results of the subword level verification to make the rejection decision for the whole keyword.

However, the aforesaid patents are used for phonetic verification, which is also used for recognizing the phonetic section of "unknown" text content, not a "known" text content used in speech corpus. Furthermore, the phonetic verification is mainly used for solving out of vocabulary (OOV) problems. But the speech synthesis technique of speech corpus needs to insure that the

recording and segmentation of every phonetic unit is correct; moreover, a target for the phonetic verification can be a word, a phase, or a sentence, which is different from a normal standard target (such as a syllable) for the speech synthesis technique of speech corpus.

Therefore, it is desirable to provide an automatic speech segmentation and verification method and related system to mitigate and/or obviate the aforementioned problems.

SUMMARY OF THE INVENTION

The object of the present invention is to provide an automatic speech segmentation and verification method and system which integrates an analysis process of a confidence measure for the segmentation verification (CMS) and a confidence measure for the syllable verification (CMV) to obtain a correct cutting position and to sieve incorrect recording material, which improves the output quality of the speech synthesis system. Another object of the present invention is to provide an automatic speech segmentation and verification method and system which automatically collects synthesis units without human effort.

To achieve these objectives, the automatic speech segmentation and verification method of the present invention first retrieves a recorded speech corpus, which corresponds to a known text script, wherein the known text script defines phonetic information with N phonetic units. Next, it segments the recorded speech corpus into N test speech unit segments referring to the phonetic information of the N phonetic units in the known text script. Then, it verifies segment confidence measures of N cutting

points of the test speech unit segments to determine if the N cutting points of the test speech unit segments are correct; and verifies phonetic confidence measures of the test speech unit segments to determine if the test speech unit segments correspond to the known text script. Finally, it determines acceptance of the phonetic unit by comparing a combination of segment reliability and the phonetic confidence measures of the test speech unit segments to a predetermined threshold value; wherein if the combined confidence measure is greater than the predetermined threshold value, the phonetic is accepted.

The automatic speech segmentation and verification system of the present invention includes a database, a speech unit segmentor, a segmental verifier, a phonetic verifier, and a speech unit inspector. The database stores a known text script and a recorded speech corpus corresponding to the known text script, and the known text script has phonetic information with N speech unit segments (N being a positive integer). The speech unit segmentor segments the recorded speech corpus into N test speech unit segments referring to the phonetic information of the known text script. The segmental verifier verifies the correctness of the cutting points of test speech unit segments to obtain a segmental confidence measure. The phonetic verifier obtains a confidence measure of syllable verification by using verification models for verifying whether the recorded speech corpus is correctly recorded. The speech unit inspector integrates the confidence measure of syllable segmentation and the confidence measure of syllable verification to determine whether the test speech unit segment is accepted

or not.

Other objects, advantages, and novel features of the invention will become more apparent from the following detailed description when taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a functional block drawing of a preferred embodiment according to the present invention;

FIG. 2 is a flowchart of the embodiment of the present invention; and

FIG. 3 is a flowchart of a segmentation process of the embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

An automatic speech segmentation and verification method and system disclosed in the present invention can be applied in establishing a speech synthesis unit of a corpus-based text-to-speech (TTS) system or any other similar application. Please refer to FIG. 1. FIG. 1 is a functional block drawing of a preferred embodiment according to the present invention. The drawing shows the preferred embodiment of the present invention as applied in a computer system 1 to perform segmenting and speech verification to a recorded speech corpus R corresponding to a known text script K. The computer system 1 comprises a database 11, a speech unit segmentor 12, a segmental verifier 13, a phonetic verifier 14, and a speech unit inspector 15 that is used for integrating the confidence measure of syllable segmentation (CMS) and the confidence measure of syllable verification (CMV) to determine whether the test speech unit segment is

accepted or not. The database is used for storing the known text script K and the recorded speech corpus R; the known text script K has phonetic information with N speech unit segments, and the recorded speech corpus R is recorded by a professional operator according to known text script K.

The following embodiment will take a Mandarin speech corpus as an example. Mandarin is a language of monosyllables, in which each Chinese character corresponds to a syllable. Therefore, it's natural and convenient to choose the syllable as the basic synthesis unit (also the speech unit segment). Phonologically, a Mandarin syllable is composed of an optional initial followed by a final. The phonemic categories of initials include nasals, liquids, stops, fricatives, and affricates. The speech unit segment can also be a signal phone, a di-phone, a root, a prefix or a word.

Please refer to FIG. 2. FIG. 2 is a flowchart of the computer system 1 of the present invention. The computer system 1 receives the recorded speech corpus R corresponding to the known text script K and stores it in the database 11 (step 201). The speech unit segmentor 12 segments the recorded speech corpus R into N test speech unit segments 121 referring to the phonetic information of the N phonetic units in the known text script K stored in the database 11(step 202) to obtain cutting point information for every test speech unit segment 121.

Please refer to FIG. 3. FIG. 3 is a flowchart of a segmentation process of the embodiment of the present invention. In order to make sure the test speech unit segments 121 maintain the phonetic features of continuous speech, the speech unit segmentor 12 uses a hidden Markov model (HMM)

to perform a segmentation alignment (step 301) to approximately segment the recorded speech corpus R into N test speech unit segments 121. Every N test speech unit segment 121 has an initial cutting point. An analysis window of the hidden Markov model has a window length of 20 ms and a window shift of 10 ms. The feature vector has 26 dimensions including 12 Mel-scale cepstral coefficients (MFCC), 12 delta-cepstral coefficients, 1 delta-log-energy, and 1 delta-delta-log-energy. The embodiment uses speaker-independent HMMs as the initial models for training the speaker-dependent HMMs.

Next, a fine adjustment is performed to every initial cutting point segmented by the hidden Markov model (step 302) to obtain a more precise cutting fine adjustment value according to some feather factors such as different unit type, speech feature data and search range. Some feather factors like a neighboring cutting point of the initial cutting point, a zero crossing rate (ZCR) of the test speech unit segments 121 and an energy of the test speech unit segments 121 can uses a window with a 5 ms length and a 1 ms shift. The energies are band-pass and high-pass signals of the test speech unit segments 121 which were obtained on a speaker-dependent band. The boundaries of Mandarin syllables, initials and finals are obtained from the state-level boundaries. As mentioned, most of the syllable boundaries are not accurate enough and need to be appropriately adjusted.

Because the segmentation is designed for a TTS system, the philosophy behind the system is how well the segmented speech can be used for concatenation. This refers to not only the accuracy of the

segmentation but also concerns how "normal" the syllable after segmentation is. Therefore, the initial cutting point and the cutting point fine adjustment value of the test speech unit segment 121 are integrated to obtain a cutting point of the test speech unit segment (step 303). However, the initial cutting point and the cutting point fine adjustment value of the test speech unit segment 121 may be integrated according to different factors, such as a different priority according to an expert's opinion, or an average opinion of disagreement among different results from multiple sources; or each cutting point fine adjustment value is provided a weighted value, and the cutting point of the test speech unit segment is a weighted average of the initial cutting point and the cutting point fine adjustment value. In addition, a duration statistics for different initials, finals, and syllable types, or an energy statistic for different initials, finals, and inter-syllable segments, or a ZCR statistic for different initials can also affect the integration.

Please refer again to FIG. 1 and FIG. 2. After obtaining the cutting point information of the test speech unit segment 121, the computer system 1 separately sends the cutting point information of the test speech unit segment 121 to the segmental verifier 13 and the phonetic verifier 14 to be verified to assure a phonetic consistency between the syllables in the recorded speech corpus R and the syllables in the known text script K. The segmental verifier 13 verifies a confidence measure for the segmentation verification (CMS) of the test speech unit segment 121 (step 203), which determines whether the cutting point of the test speech unit segment 121 is

correct according to various phonetic statistical factors to determine a boundary of the test speech unit segment 121. The phonetic verifier 14 verifies a confidence measure for the syllable verification (CMV) of the test speech unit segment 121 (step 204), which determines whether the test speech unit segment 121 corresponds to the known text script K. The above-mentioned step 203 and step 204 are preferably performed at the same time or, their performance sequence may be switched.

The confidence measure for the segmentation verification can be defined as:

$$CMS = \max\left(1 - h(D) - \sum_{s,f} g(c(s), f(s)), 0\right),$$

and the item h(D) is the disagreement function, which can be defined as $h(D) = K\left(\sum_i w_i \left| d_i - \bar{d} \right|\right)$ (or $h(D) = K\left(\sum_i w_i (d_i - \bar{d})^2\right)$), where D is a vector of multiple expert decisions of the cutting point of the test speech unit segment 121, d_i is the cutting point, and $\bar{d} = p(D)$ is a final decision of the cutting point determined by the priority, average value or weighted value; K(x) is a monotonically increasing function that maps a non-negative variable x into a value between 0 and 1. The function h(D) is used to verify an inconsistency among multiple decision points in a multiple expert system. A high inconsistency means a low confidence measure for the syllable verification g(c(s), f(s)) is a cost function value between a cost function ranging from 0 to 1, s is a segment, c(s) is a type category of the segment s and, f(s) are acoustic features of the segment (including a zero crossing rate,

duration, energy, and periodicity) of the segment, respectively. Therefore, the confidence measure for the syllable verification CMS in this embodiment is between 0 (a lowest confidence measure) and 1 (a highest confidence measure). During the calculation, if there is any unconfident measure the CMS is reduced down to 0. The unconfident measure can be a distinction of the cutting position. For example, if the segmented test speech unit segment 121 is expected to be fricative which will has a higher ZCR, but is actually a silence, and an actual ZCR is not as high as it was expected, then this will cause a calculated cost to become larger and reduce the confidence measure.

The phonetic verifier 14 performs automatic syllable verification according to a discriminative utterance verification for non-keyword rejection in the field of speech recognition. The phonetic verifier 14 uses a syllable model and an anti-model for each syllable type for every syllable to combine a phonetic verification model. The syllable model is trained for verifying the possibility of the test speech unit segment 121 meeting the target syllable type. The anti-model, on the other hand, is trained for verifying the possibility of the test speech unit segment 121 not meeting the target syllable type. Since the method of training the syllable model and the anti-model is well known to those familiar with this field, no further description is required. Consequently, the phonetic verifier 14 can obtain:

$$CMV = \min\{LLR_I, LLR_F, 0\},$$

where, $\begin{cases} LLR_I = \log P(X_I \mid H_0) - \log P(X_I \mid H_1) \\ LLR_F = \log P(X_F \mid H_0) - \log P(X_F \mid H_1) \end{cases}$, X_I is the initial segment of the test speech unit segment, X_F is the final segment of the test speech unit segment 121, H_0 is null hypothesis of the test speech unit segment 121

recorded correctly, H_1 is alternative hypothesis of the test speech unit segment 121 recorded incorrectly, and LLR is a log likelihood ratio.

The segmental verifier 13 in this embodiment can not only verify the cutting position but also partly verify a confidence measure for phonetics. When the recorded speech corpus R is incorrect, the related feature factor will be wrong. Furthermore, the phonetic verifier 14 can not only verify phonetic content but also the cutting position. Because, when the cutting position is wrong, the phonetic verification will have a larger mistake, which reduce the confidence measure for the syllable verification (CMV).

Finally, the speech unit inspector 15 integrates the confidence measure of syllable segmentation (CMS) and the confidence measure of syllable verification (CMV) and compare to a predetermined threshold value (step 205) to determine whether to accept the test speech unit segment 121 (step 206). The speech unit inspector 15 can choose an early decision method or a late decision method to compare the confidence measure of syllable segmentation (CMS) and the confidence measure of syllable verification (CMV) of the test speech unit segment 121. The early decision method has two different methods; the first method separately compares the confidence measure of syllable segmentation (CMS) and the confidence measure of syllable verification (CMV) to the threshold value, and only accepts the test speech unit segment 121 in the speech corpus if both of the confidence measures are larger than the threshold value; for the second method, if one of the confidence measures is larger than the threshold value, the test speech unit segment 121 is accepted. The late decision method standardizes the confidence measure of syllable segmentation (CMS) and the confidence measure of syllable verification (CMV), and gives different weighted values

to different confidence measures to calculate a signal confidence measure that is compared to the threshold value for the decision.

The present invention provides automatic speech segmentation, segmentation integration and a novel phonetic verification process to obtain correct cutting positions and to detect problematic speech segments, which largely reduces human labor in construction, and increases the quality of the speech corpus.

Although the present invention has been explained in relation to its preferred embodiment, it is to be understood that many other possible modifications and variations can be made without departing from the spirit and scope of the invention as hereinafter claimed.